# Data from Non-standard Varieties

Nerbonne, John

*Published in:*
Proceedings of the 13th Conference on Natural Language Processing

# Data from Non-standard Varieties

**John Nerbonne**
Dept. Informatiekunde
Rijksuniversiteit Groningen
&
Germanistische Linguistik
Albert-Ludwigs-Universität Freiburg
j.nerbonne@rug.nl

## Abstract

The most important reasons for examining "non-standard data" with CL methods are the facts that this data represents a great deal of language behavior and that it serves as an object of scientific study in linguistics as a whole. This is true of the syntax of non-native second-language learners, the accents of non-native speakers, and the vocabularies of different dialect speakers.

Computational linguists have a good deal to offer to the various subfields of linguistics studying non-standard data. By automating steps in analysis we make the analyses replicable and also modifiable, we improve opportunities for comparing similar analyses, and perhaps most importantly, we enable the analyses of large amounts of data, providing more comprehensive views.

The data itself can be tricky to work with, however, as scientists in other fields are often specialized in a single language or language pair, which means that their data will not be varied enough to support all the research questions one would like to ask, e.g., the question of the generality of the techniques for a particular purpose. In other cases, the data simply won't have been collected with an eye to answering some interesting questions, which may mean that important parameters haven't been recorded. Finally, we note that non-automated analyses do not impose expectations that data be commensurate to the same strict degree (as automated ones), meaning that surprises can be in store even in well-studied data sets. This paper provides some concrete examples and discussion of these potential pitfalls.

One can protect oneself from some of these risks by seeking collaboration with domain experts, which is to be recommended in any case, as a way of making the work richer and better informed. Further, it makes sense to approach novels sorts of data — and even novel sources of data of a sort one suspects is familiar — with a broad range of potential research questions. There is an awful lot of interesting work still to be done!

## 1 Introduction

The theme of this year's KONVENS is non-standard data, and it's a great choice as computational linguistics (CL) ventures into areas of linguistics it's traditionally shied away from! I interpret the shyness incidentally not as a lack of CL interest in areas such as spoken language, historical data, second language learning, etc. (topics mentioned in the call for papers), nor as disregard for non-standard varieties, and certainly not as indifference to unedited prose in general, but rather as a wish to concentrate on honing technique and a wish to obtain results that allow interpretation with a focus on technique. From those points of view it makes sense to limit other parameters from varying too much. But I also agree emphatically that the time is ripe to widen CL's purview to include language from these other areas.

There are several reasons why we as computational linguists should work more with less standard data. First, most language is produced without any editing, and therefore without any effort to put it into a standard form. If we're going to deal with language of a wide variety of sorts, it will be difficult to avoid non-standard data. Second, there are important contributions CL is poised to make and which are simply required to make progress in this area. Some recent papers that illustrate this are Eisenstein, O'Connor, Smith, and Xing (2014) and Jurafsky, Chahuneau, Routledge, and Smith

(2014). Nguyen, Doğruöz, Rosé, and de Jong (Accepted to appear) provides a survey of CL work related to sociolinguistics, a large part of which involves the analysis of social media, usually in rather spontaneous, i.e. non-standard form.

I want to contribute to this theme by relating some of my experience with working with non-standard data, in particular data from non-standard varieties — dialects, "regiolects" (intermediate between dialects and stand languages, see Auer and Hinskens (1996)), and language in situations where there is contact (second language learner situations). After relating some of this experience, I'll close with some reflection on these. My intent is to be encouraging, but I'll note pitfalls as well as opportunities. There's every reason to be keen, but also to be cautious.

## 2 Contact syntax, aggregate distance, and detecting differences

Some colleagues in Linguistics at the University of Oulu were eager to collaborate on the *Finnish Australian English Corpus* (Watson, 1996). They'd worked on the corpus before, but never using language technology. The corpus consists of transcriptions of conversations held with Finnish emigrants to Australia. The 60-member group we'll focus on were adults on emigrating in the 1960s (around 30 yr. old), and they were interviewed after 30 years in Australia. They had working-class backgrounds, and "very few could speak any English at all on arrival to Australia" (Watson, 1996, p.45). The English was indeed quite rough, as expected, and this of course posed the technical challenge. To get a flavor of this consider the following excerpt from the corpus, elicited by asking participants to describe what a soldier would need to do to complete an assault course they were shown in a sketch:

> The soldier first have to go climb, climb to tree. Then uh, I don't know how they call that but uh, I, I call um, walkin' by hands, hangin' by hands or walkin' hands to other tree, come down to ground, walkin', um, uh, not walkin' but climbin' over brick wall, come dine..., do..., down other side, then have to go to ground by knees, goin' under some or, or whatever it is, climbin' up by ladders to other bick..., brick wall and jump down to ground on other side + um, there is, then have to go tunnel, maked from brick, come out on other end and ju..., jump to river, swim cross to finish line.

### 2.1 Theoretical goals

There is and was scientific consensus that one should expected to find Finnish-like elements in the speech of these emigrants (Opas-Hänninen, Hirvonen, Juuso, & Lauttamus, 2005), but I felt especially challenged first by a remark by the great theoretician on language contact, Uriel Weinreich:

> No easy way of measuring or characterizing the total impact of one language on another in the speech of bilinguals has been, or probably can be devised. The only possible procedure is to describe the various forms of interference and to tabulate their frequency (Weinreich, 1968, p.63)

Second, Ellis (1994), De Bot, Lowie, and Verspoor (2005) and other theorists of second-language acquisition have emphasized that it is not enough to catalog the "errors" of second-language users, because non-native speech often differs not in errors, but rather by overusing and by underusing specific linguistic items, where easier elements are typically overused and tougher ones underused. So the initial goal was to develop an aggregate measure of syntactic difference that was sensitive to overuse and underuse.

We settled on looking at part-of-speech (POS) tags, focusing on the frequency distributions of trigrams of POS-tags. We deliberately did not include lexical information in order to focus on syntax, and we decided to use trigrams in order to make the measure sensitive to context. Of course we were aware that looking at ordered sequences of syntactic categories might not be a general solution, but we were looking at English, where order is quite important, and we were interested in language behavior, not language competence.[1] By examining frequencies, we automatically gauged overuse and underuse, and by examining the entire distribution of POS-tag sequences we could claim to be contributing to Weinreich's goal of providing information on the total impact of the first language on the second (albeit only for syntax). We set our initial goals quite high.

[1] Sanders (2007) extended the work under description by examining leaf-path ancestors in parse trees.

### 2.2 Results

But would it work technically? We trained Thorsten Brant's TnT tagger (Brants, 2000) on the British part of the ICE corpus using the 270-element TOSCA-ICE tag set (Nelson, Wallis, & Aarts, 2002). We were naturally concerned with tagging accuracy, so we manually evaluated tagging accuracy on 1,000 randomly chosen words. The tagger was correct 81.2% of the time for single tags dropping to 56.1% for trigrams (Brants's tagger is 96.7% accurate when applied to the Penn Treebank) . See Wiersma, Nerbonne, and Lauttamus (2011). We also experimented with a smaller tag set, which naturally improved performance, but we decided to use the larger set for its more sensitive reflection of syntax.

We also tagged a corpus of speakers who had emigrated at 17 years of age and younger, because the material was most commensurable to the transcripts of the older emigrants. The speech of the younger emigrants was native-like, and we used this to identify particularly deviant POS trigrams.

We ignored very infrequent POS-trigrams (nearly 40,000 trigrams with frequency less than five in either corpus) so as not to be misled by what might be coincidence, and then compared the relative frequencies in the two corpora under comparison. Wherever the relative frequencies differ a good deal, we suspect that we are seeing contact-induced effects.

With respect to developing a syntactic distance measure, a rigorous validation would have to compare several data sets, ideally involving different target and different source languages, as well as several degrees of non-nativelike syntactic behavior.[2] So the best we can say on this score is that we've introduced a technique, but not that we've shown it to be probative, and certainly not for a range of languages and different degrees of contact-induced "contamination".

Close collaboration with the domain expert, Timo Lauttamus, was absolutely essential in applying this work to the question of detecting differences automatically. He examined a random sample of 137 of the 300 most divergent POS-trigrams and showed that most of them — all but 24 — were interpretable as the result of second-language disfluencies and a Finnish "substrate" in the emigrants' English. These include problems with (in)definite articles (Finnish has none), with the copula *be* (missing in Finnish), with the expletive *there* (likewise missing), and difficulties with contractions and auxiliary verb sequences, both of which led to underuse. See Lauttamus, Nerbonne, and Wiersma (2007) for a detailed presentation. From the point of view of identifying deviant syntax, the work was successful.

[2]For the sake of completeness I'll add that we *could* test for overall differences — in line with Weinreich's goal — by applying a permutation test to the table with two varieties and 8,300 instantiated POS-trigrams. This involved a tricky normalization. Di Buccio, Nunzio, and Silvello (2014) have suggested using vector space techniques to compare the trigram frequencies, and this seems more straightforward.

### 2.3 Reflection

It should be clear that we cannot claim to have solved the problem of measuring overall syntactic differences in varieties. We developed a measure and showed that it could be put to good exploratory use, but we certainly do not claim to have validated it rigorously. We just showed that the software helped in looking for differences in language use — in spite of the fact that the data was certainly noisy and the computational tool suboptimal.

Collaboration with Lauttamus, an expert on the English Finns acquire naturally, was essential to the success of the project, as was the fact that we eschewed a narrow focus on developing a measure of aggregate syntactic difference. I think we pushed the envelope a bit on that score, but, as I've emphasized, it would be rash to claim success on that point. It was essential that we aimed rather broadly in dealing with this data.

## 3 Accents and a caution on theory

There is a well-established line of research in which CL techniques are applied in dialectology, and Nerbonne (2009) motivates this theoretically. For the most part, this line of research has applies edit-distance measures (Kruskal, 1983) to phonetic transcriptions, and the work has established itself at least in areas where transcriptions are the primary recordings of pronunciations (most data collections more than about fifty years old). We have experimented with various modifications to the basic edit distance algorithm, where Heeringa, Kleiweg, Gooskens, and Nerbonne (2006) gives a flavor of the range of these modifications we've experimented with. Currently we prefer a version in which segment distances are weighted by the (inverse) frequency of their chance of corresponding in alignments. Because we gauge this frequency in-

formation theoretically, using pointwise mutual information, we refer to this as PMI-LEVENSHTEIN (Wieling, Margaretha, & Nerbonne, 2012).

But it has always been clear that reliable procedures for assaying the degree of difference in pronunciation would be useful for other reasons. Kondrak and Dorr (2006) demonstrate that this sort of procedure, applied to candidate drug names against the background of a data set of existing names, can identify potentially confusing name candidates, a circumstance that has been shown to have occasionally fatal consequences. In information retrieval, it is often difficult to find references to people whose names are normally spelled in a different writing system (Nabende, Tiedemann, & Nerbonne, 2010), such as Cyrillic, Arabic, Urdu or Japanese. One example of such a name is 'Musharraf', which sometimes occurs as 'Musharrav', 'Musharaf', etc. While there are often established conventions for TRANSLITERATION, few writers obey these, so a common technique is to attempt to identify alternative spellings that are likely to have the same pronunciation. This makes procedures for measure pronunciation differences useful in this context as well. Yet another, third area of application is in the diagnosis of speech problems, and Sanders and Chin (2009) have indeed applied an edit distance measure of the speech of cochlear implant bearers with some success.

### 3.1 Accents

It had occurred to me and to others that measuring how strong foreign accents are might be a fourth area of linguistics where a measure of pronunciation differences might be of interest, in particular to researchers in second-language learning. So I was very pleased when one of my collaborators, Martijn Wieling, noticed the Speech Accent Archive at George Mason University (Weinberger & Kunath, 2011). It contained then the recordings of over 800 non-native speakers of American English together with their phonetic transcriptions. By organizing a web-based judgment task[3] we were able to validate the PMI-based edit distance for this slightly different task — that of judging how non-native a speech sample sounded. The computational measures correlate very strongly ($r = 0.81$) with the judgment of native speakers with respect to how native-like the recorded passages were (Wieling, Bloem, Mignella, Timmermeister, & Nerbonne, 2014).[4] So the effort of moving into a new field led to a valuable new validation of technique, and this is worthwhile!

We were also able to explore the scientific issues a bit, investigating factors influencing the quality of the non-native accent, both the age at which English was learned and also the number of years resident in an English speaking country. These "insights" are almost proverbial — amounting to "the early bird catches the worm" and "practice make perfect", so we certainly don't claim any scientific breakthrough here, but our sample was large enough to let us catch a non-linear interaction between the two sorts of influences. As Figure 1 demonstrates, the two factors interact in a complex way. The "contour lines" on the regression surface are not evenly spaced as one moves up the age of learning onset; instead, lines are further and further away from one another, meaning that it becomes harder and harder to compensate for a late start with a long residence. I think we are the first to show this (Wieling, Bloem, Baayen, & Nerbonne, 2014).

### 3.2 Overreaching theoretically

So far, my report on this foray into a new sort of data makes it sound like an unqualified success, but there is more to tell. In a further step we tried to apply our "insights" to illuminate a famous issue in language acquisition and cognitive science, namely the CRITICAL AGE HYPOTHESIS. The idea is straightforward. We let the the distance of the learner's speech from the native pronunciation stand proxy for the success of language acquisition in general, and take the age of learning onset at face value. We can then plot the distance of the learner's speech from the native pronunciation as a function of the age of learning onset to get an idea if whether the decline in ultimate attainment is smooth, or whether there is a point — sometime before eighteen years or so — where ability sharply decreases. There's a nice paper by Jan Vanhove reminding us that PIECEWISE REGRESSION is the right technique to apply statistically (Vanhove, 2013), and Figure 2 shows the result of applying piecewise regression to the accent data.

Figure 2 breaks the data down into speakers of Indo-European languages (IE) and non-IE languages, which was not part of an initial hypothesis

[3] We're grateful to Mark Liberman for announcing this on *Language Log*, which is why so many subjects joined in.

[4] The native speaker judges only agreed with each other to a slightly greater degree ($r = 0.84$)

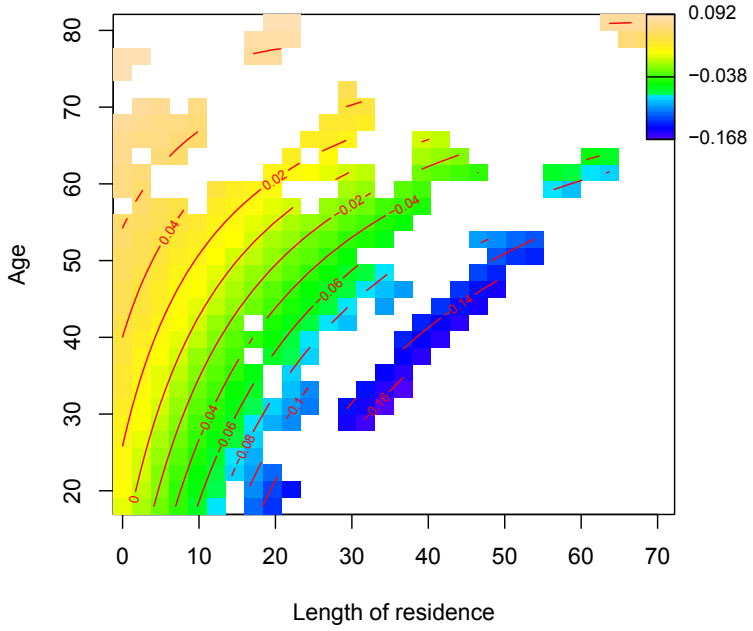


Figure 1: Accent quality is shown by the color, ranging from dark blue (quite native-like) to light yellow (distinctly foreign accent). Note that, in general, a long residence leads to better accents (darker blue) as does an early age at which English was learned. White areas indicate combinations of parameters for which little data was available.

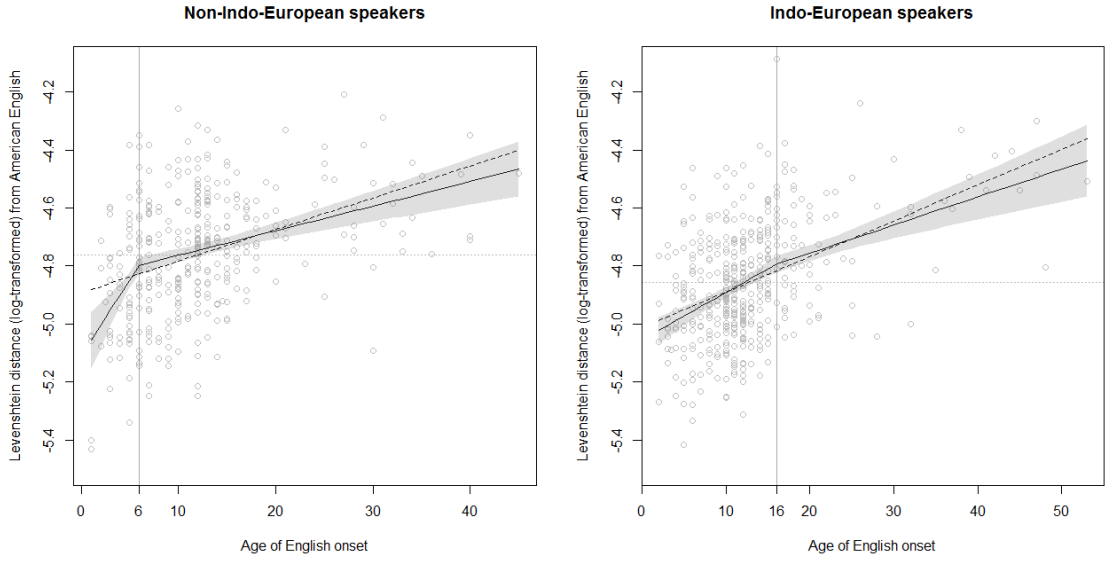


Figure 2: Accent quality (non-nativelikeness) deteriorates monotonically as a function of the age at which English was first learned. Moreover, there appears to be a sharp break around six years of age for speakers of non-IE languages (left), which would be compatible with the critical age hypothesis.

we brought to the data, so clearly **not** a hypothesis we might claim to confirm based on the analysis, but it's quite intriguing, as it suggests that the native language of learners might be a confound in studies of ultimate attainment of second-language learners. Speaker of languages related to English deteriorate throughout their lifetimes in their approximation to English pronunciation, but the deterioration is fairly constant. In contrast, the non-IE speaker's decline changes abruptly, even though, curiously, the rate of decline decreases after a critical point.

The referees varied in their reactions. One was definitely positive about the novelty of the finding and correctly chided us for failing to acknowledge potential biases in the data set (selection!), which the others likewise saw, but we were chastised to the point of emphatic rejection for not being *au courant* in the literature on second language learning vis-à-vis the critical age hypothesis. We'd read up on what we could, but there's an enormous literature, and it is difficult to get a sense of all that specialists hold dear. We definitely failed to distinguish IMPLICIT learners – those with no formal language training – from EXPLICIT learners, while the field has turned to seeing only implicit learners as interesting. In fact, however, there is no good way to operationalize this distinction in the Speech Accent Archive (Weinberger & Kunath, 2011) — those compiling the data set simply didn't include this information. So this aspect of the work simply failed to show what we originally claimed.

### 3.3 Reflection

Just as in the experiences with the non-native syntax, this line of research achieved some success, and for that it was again crucial that we had aimed broadly — both at validation of PMI-Levenshtein as a measure of pronunciation difference and at characterizing the role the age of learning onset and the length of residence plays. We failed to contribute to the discussion on the critical age for language acquisition due to our not knowing the literature sufficiently.

In retrospect we became convinced that there was no way to use the data to say much of anything about the critical age hypothesis, and it would have been prudent to seek collaboration with a language acquisition specialist before developing that aspect of the work. So the point for computational linguists interested in non-standard data is just that there is often a body of theory and and a literature that one simply has to command sufficiently in order to contribute.

## 4 Lexical variants and incommensurate data

Working with non-standard data entails surprises, at least occasionally. Working with standard data — say the Penn Treebank, the BNC or CELEX — means building on the work of others and (normally) relying on the intelligent choices of predecessors. Leaving this well-trodden path means that one occasionally has to think through the whole process of what it means to draw inferences with respect to a given hypothesis based on an unfamiliar data source. It often entails working with data that has previously only been analyzed manually and perhaps only examined for key features, so that there may be no experience of automatic processing, which means in turn that some problems — including missing data, confounds and unexpected distributions — may arise for the first time.

### 4.1 Dialect variation in vocabulary

It is interesting to examine the degree to which different linguistic levels correlate in their geographic distribution, e.g., pronunciation, lexical choice and syntax (Spruit, Heeringa, & Nerbonne, 2009), so we have looked at lexical and syntactic variation as well the pronunciation variation that we've mostly concentrated on. One such study involved the *Linguistic Atlas of Middle and South Atlantic States* (LAMSAS, Kretzschmar Jr. (1993)). In comparison to applying appropriate edit distance measures (to non-standard transcriptions), the vocabulary task sounded simple. Varieties should count as the same to the degree that they use the same words in response to fieldworkers' questions.

Looking at the data convinced us of two things, first, that simple string identity was likely to be too rough a measure to be useful. See Table 1, and see `http://www.let.rug.nl/~kleiweg/lamsas/overview/lex.txt` for a complete listing of all the responses in the data set.

Nerbonne and Kleiweg (2007) present a range of techniques that have been proposed for detecting similarity in lexical data, including approaches that use Porter stemming, edit distance, and inverse frequency weighting (Goebl, 1984) in (five) various combinations. This paper is written in the usual style of computational linguistics (CL), where several techniques are compared with respect to the

*clearing up (435), clearing off, clearing, fairing off, clear up (50), fairing up, clear off, cleared up, fair off, clearing away (28), cleared off, breaking off, faired off, breaking away, fair up (18), break off, breaking, going to clear up, clear, fairing (9), ..., clouds is breaking (3), ..., ceasing, changing, fair, ..., held up, is broke, weather's going to break (2), a Dutchman's britches (1), ..., a-fairing, ..., a settling off, ..., blow off, blue sky enough to, ..., brightening, ..., make a Dutchman's pants, ..., moderating, ..., slacked up, ...*

Table 1: Selection of responses to the question "If the sun comes out after a rain, you say the weather is doing what?" in decreasing order of frequency. 1516 response tokens, including 81 singletons (hapax legomena).

performance on an object measure. In the interest of space, I will not not repeat the presentation here, but I will note that it demonstrates the usefulness of CL techniques on non-standard data.

Second, we noted that field workers had often recorded multiple responses. Since this gives a flavor of working with non-standard data, I'll summarize the treatment here. For example, there were 1516 responses to the question of how to describe weather when rain was giving way to fairer skies — coming from only 1162 informants. Given that the data included multiple responses, we had to develop a generalization of the simple identity criterion for scoring responses. After all, the distance between $\{a,b\}$ and $\{a\}$ ought to be larger than the distance between $\{a\}$ and itself but smaller that the distance between $\{a\}$ and $\{b\}$:

$$d(\{a\},\{a\}) < d(\{a\},\{a,b\}) < d(\{a\},\{b\})$$

One might think of simply using the mean distance of the cross product between sets $A$ and $B$ of responses, but would make the distance between the $\{a,b\}$ and $\{b,a\}$ non-zero, so we developed a measure that is slightly more abstract, arriving at the following definition:

$$d(A,B) \doteq \frac{1}{|C|} \text{ Min } d(C), \qquad \text{where } C \text{ covers } A \times B$$

We stipulate that a set of ordered pairs $C$ COVERS $A \times B$ as long as every element in $A$ occurs as the first element of some pair in $C$ and likewise every element of $B$ occurs as a second element in a pair in $C$. $d(C)$ is just the sum of the distances in the set of ordered pairs. Note that this definition has the consequence that $d(\{a,b\},\{b,a\}) = d(\{a,b\},\{b,a\}) = 0$ The minimum cost cover in this case is $\{< a,a >, < b,b >\}$, whether the distances sum to 0.

### 4.2 Surprising preliminary analyses

After working out potential solutions to these two issues, we proceeded to first analyses, and were surprised when we clustered the aggregate lexical distances to obtain the result on the left in Figure 3, which doesn't correspond in the least to anything we'd read on American dialect areas! Given how instable clustering sometimes can be, we verified the analysis by applying multi-dimensional scaling (MDS, Nerbonne, Heeringa, and Kleiweg (1999)) to the aggregate distance table, but the impossible division cannot be blamed on clustering.[5] As the middle map shows in Figure 3 shows, the picture is even more incoherent when we include larger numbers of clusters.

After a good deal of exploration in the LAMSAS, including analysis of the various questionnaires used the years in which interviews were held, Peter Kleiweg noticed that the field workers differed enormously in the number of responses they recorded. Figure 4 shows that while Lowman was remarkably consistent in recording about the same number of responses in each interview, the other field workers were much less consistent. We also considered trying to use only the first response provided, but it wasn't clear that the first response provided represented the preferred response of the informant. The fact that the fieldworkers had collected essentially incommensurable sets of responses hadn't handicapped earlier, manual work with the data set, but I think that we were the first to point out that the discrepancies existed. Fortunately, Lowman's consistency meant that we could conduct and publish an analysis on a substantial subset of the LAMSAS data (Nerbonne & Kleiweg, 2003). Figure 5 show the areal division arising from the treatment sketched here; it corresponds

[5] Leinonen, Çöltekin, and Nerbonne (2016) present an MDS check on clustering results into the *Gabmap* web application for the analysis of language variation.

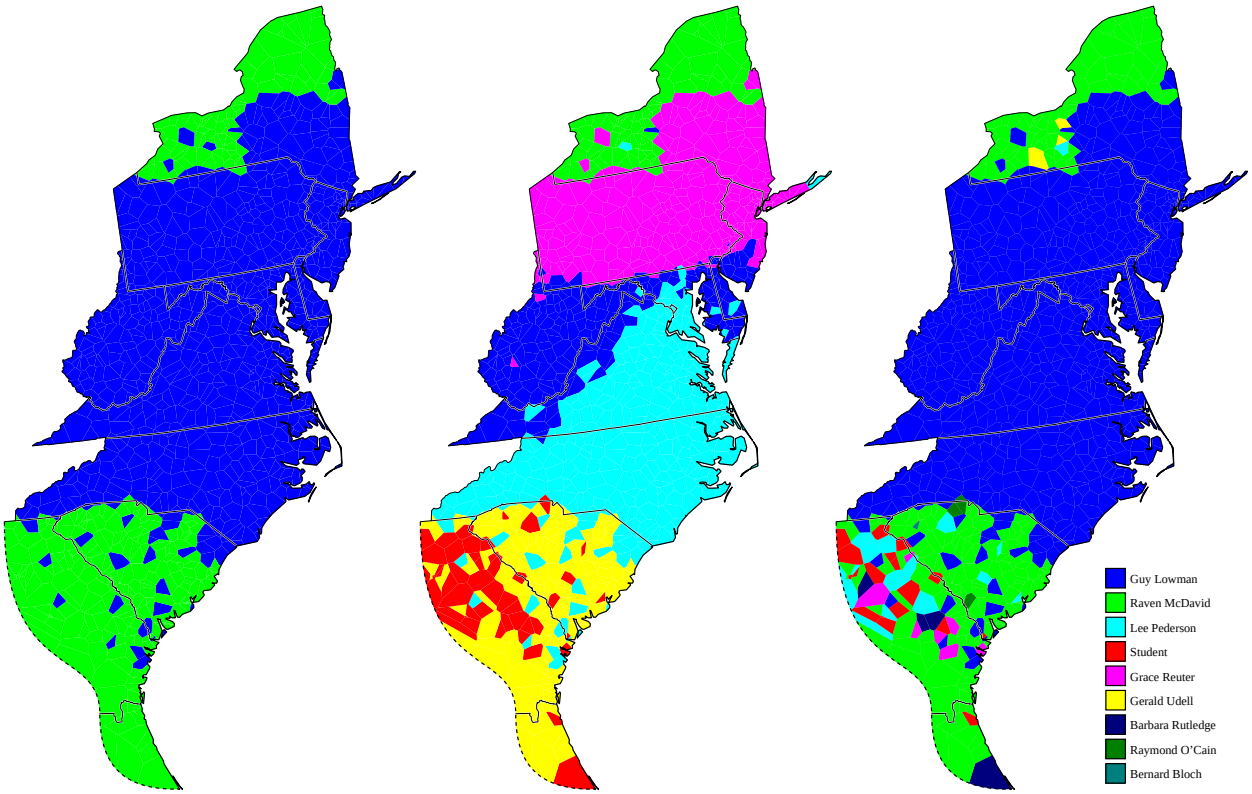


Figure 3: Preliminary results of clustering based on lexical choice in LAMSAS, where the legend on the right, showing the fieldworkers responsible for the data collection, provides an interpretation only for the rightmost map — i.e., where the fieldworker collected data. Presented at *Methods in Dialectology X*, Joensuu, but not included in the black and white publication.

closely to a controversial division originally proposed by Kurath (1949).

It turns out, by the way, that we were able to correct for the differences between the two fieldworkers at an aggregate level, essentially using standardized scores determined for each fieldworker in turn ($z$-scores), which we then used for comparisons. Wieling and Nerbonne (2011) use a correction on transcription practices to deal with a similar problem involving comparison in a data set where two transcription teams disagreed.

### 4.3 Reflections

So the degree of success in the work on lexical overlap among the LAMSAS sites is mixed. We were able to compare different standard CL techniques — stemming, and edit distance – as well as inverse frequency weighting (appealed to in particular as a means of detecting historical affinity) in order to make sense of a difficult data set. Further, we were able to extend the normal comparison of categorical data (same vs. different) to situations in which multiple responses are found.

But we were nonetheless taken aback by how incommensurable the data was with respect to the different field workers. Using what are common CL techniques (with the extensions mentioned) enabled a lexical analysis of the full set of responses for about 70% of the data, but the differences in the number of responses collected per field was never settled satisfactorily (*pace* the remarks above). It was a lucky coincidence that one fieldworker had collected 70% of the data, that he was very consistent in the number of responses he elicited, and that the area he worked in was geographically coherent. This meant that an analysis of his data alone was worthwhile.

Overall the exercise was successful, but it certainly illustrates how easily one can be surprised by non-standard data.

## 5 Final reflections

The most important reasons for examining non-standard data with CL methods are the fact that non-standard data represents a great deal of language behavior, and that it serves as the object of scientific study in linguistics as a whole. This is true of the syntax of non-native second-language learners, the accents of non-native speakers, and the vocabularies of different dialect speakers.

Computational linguists have a good deal to offer to the various subfields of linguistics studying non-standard data. By automating steps in analysis we make the analyses replicable (and modifiable), we improve opportunities for comparing similar analyses, and perhaps most importantly, we enable the analyses of large sets of data, providing more comprehensive views.

The data itself can be tricky to work with, however, as scientists in other fields are often specialized in a single language or language pair, as we saw in the case of the work on the syntax of Finnish emigrés to Australia, and this means that the data will not be varied enough to support all the research questions one would like to ask — in this case the question of the generality and validity of the techniques for a range of cases. In other sub-disciplines, the data simply won't have been collected with an eye to answering some interesting questions, as we saw in the case of the foreign accents, where, we hasten to add, the restriction might have been obvious to researchers who had familiarized themselves with the theoretical discussion beforehand. Finally, we note that non-automated analyses do not impose expectations that data be commensurate to the same strict degree (as automated ones), meaning that surprises can be in store even in data sets that are respected as standards in the field. The LAMSAS data provides an example of this.

One can protect oneself from some of these risks by seeking collaboration with domain experts, which is to be recommended in any case, as a way of making the work richer and better informed. Further, it makes sense to approach novels sorts of data — and even novel sources of data of a sort one suspects is familiar — with a broad range of potential research questions.

There is an awful lot of interesting work still to be done!

## References

Auer, P., & Hinskens, F. (1996). The convergence and divergence of dialects in Europe. New and not so new developments in an old area. *Sociolinguistica*, *10*, 1–30.

Brants, T. (2000). TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (pp. 224–231).

De Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition: An advanced resource book*. Psychology Press.

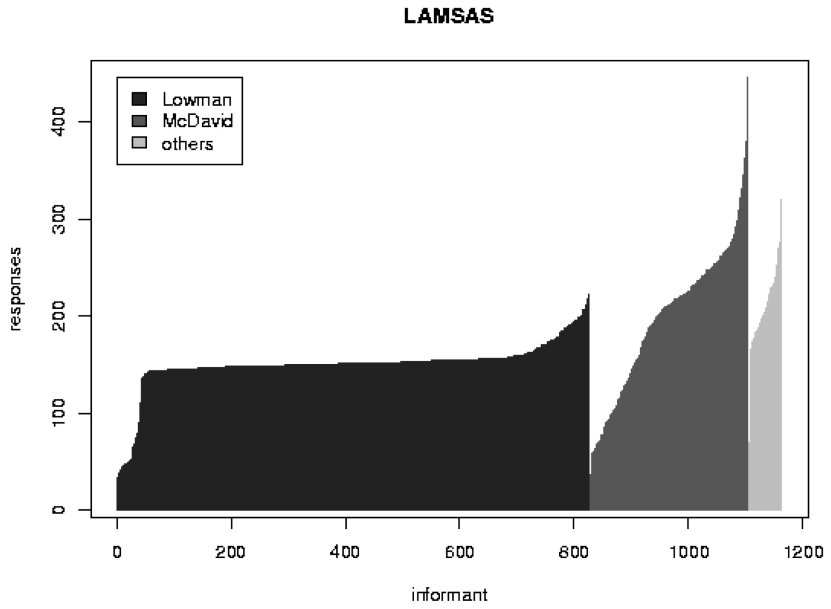


Figure 4: The number of responses per interview broken down by fieldworker. Lowman's interviews were remarkably consistent, allowing comparative interpretations, while others were not. From Nerbonne and Kleiweg (2003)

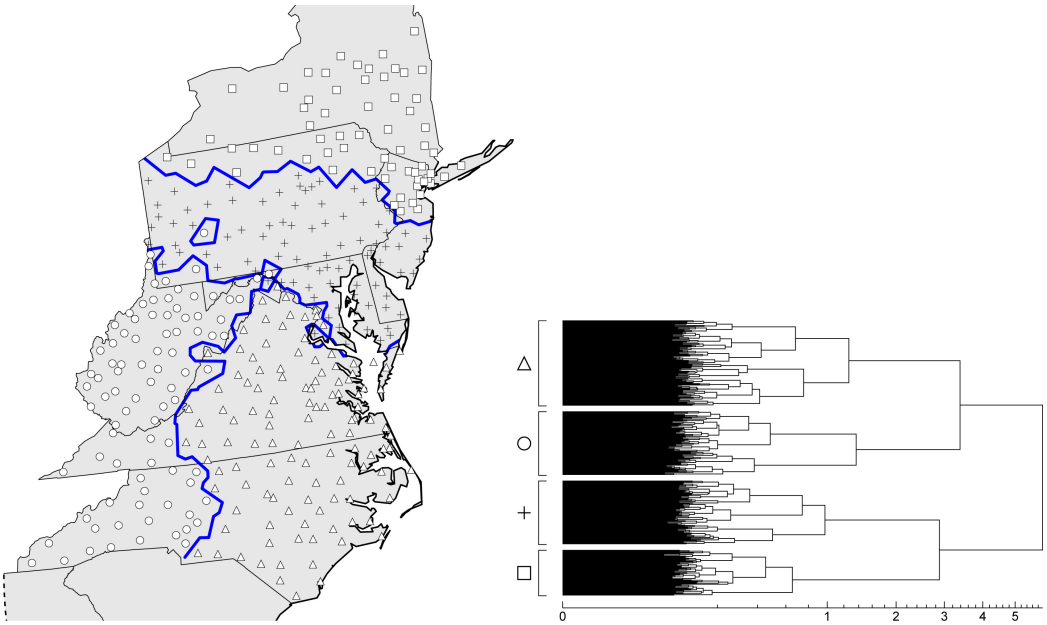


Figure 5: The final analysis of Lowman's lexical data, which, incidentally jibes well with Kurath's division. From Nerbonne and Kleiweg (2003)

Di Buccio, E., Nunzio, G. M. D., & Silvello, G. (2014, May). A vector space model for syntactic distances between dialects. In N. Calzolari et al. (Ed.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland: European Language Resources Association (ELRA).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS ONE*, *9*(11), e113114.

Ellis, R. (1994). *The study of second language acquisition*. Oxford University.

Goebl, H. (1984). Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol. *Tübingen: Max Niemeyer*.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In *Proceedings of the Workshop on Linguistic Distances* (pp. 51–62).

Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, *19*(4). doi: http://dx.doi.org/10.5210/fm.v19i4.4944

Kondrak, G., & Dorr, B. (2006). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29–42.

Kretzschmar Jr., W. A. (1993). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. University of Chicago Press.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, *25*(2), 201–237.

Kurath, H. (1949). *A word geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Lauttamus, T., Nerbonne, J., & Wiersma, W. (2007). Detecting syntactic contamination in emigrants: the English of Finnish Australians. *SKY Journal of Linguistics*, *20*, 273–307.

Leinonen, T., Çöltekin, Ç., & Nerbonne, J. (2016). Using Gabmap. *Lingua*, *178*, 71–83.

Nabende, P., Tiedemann, J., & Nerbonne, J. (2010). Pair hidden Markov models for named entity matching. In T. Sobh (Ed.), *Innovations and advances in computer sciences and engineering* (pp. 497–502). Springer.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins Publishing.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, *3*(1), 175–198.

Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Edit distance and dialect proximity. In D. Sankoff & J. Kruskal (Eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison* (2nd ed., p. i-iv). Stanford: CSLI.

Nerbonne, J., & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, *37*(3), 339–357.

Nerbonne, J., & Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, *14*(2-3), 148–166.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (Accepted to appear). Computational sociolinguistics: A survey. *Computational Linguistics*. Retrieved from `arXivpreprintarXiv:1508.07544`

Opas-Hänninen, L. L., Hirvonen, P., Juuso, I., & Lauttamus, T. (2005). Happen I not talking good English: The progressive aspect in the English of Finnish Australians. In *Methods XII: Twelfth international conference on methods in dialectology* (pp. 1–5).

Sanders, N. C. (2007). Measuring syntactic difference in British English. In *Proceedings of the ACL 2007 Student Research Workshop* (pp. 1–6). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P07-3001`

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, *16*(1), 96–114.

Spruit, M. R., Heeringa, W., & Nerbonne, J. (2009). Associations among linguistic levels. *Lingua*, *119*(11), 1624–1642.

Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLoS ONE*, *8*(7), e69172.

Watson, G. J. (1996). The Finnish-Australian

English corpus. *ICAME Journal*, *20*, 41–70.

Weinberger, S. H., & Kunath, S. A. (2011). The Speech Accent Archive: Towards a typology of English accents. In J. Newman, R. H. Baayen, & S. Rice (Eds.), *Corpus-based studies in language use, language learning, and language documentation* (pp. 265–281). Amsterdam: Rodopi.

Weinreich, U. (1968). *Languages in contact*. The Hague: Mouton.

Wieling, M., Bloem, J., Baayen, R. H., & Nerbonne, J. (2014). Determinants of English accents. In J. Wahle, M. Köllner, H. Baayen, G. Jäger, & T. Baayen-Oudshoorn (Eds.), *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics.* (Data package in *The Mind Research Repository*, Potsdam) doi: http://dx.doi.org/10.15496/publikation-8628

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. (2014). Automatically measuring the strength of foreign accents in English. *Language Dynamics and Change*, *4*(2), 253–269.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307–314.

Wieling, M., & Nerbonne, J. (2011). Measuring linguistic variation commensurably. *Dialectologia: revista electrònica*, *Special issue II*, 141–162.

Wiersma, W., Nerbonne, J., & Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, *26*(1), 107–124. doi: 10.1093/llc/fqq017